FULL LEGAL NAME	LOCATION (COUNTRY)	EMAIL ADDRESS	MARK X FOR ANY NON-CONTRIBUTING MEMBER
Aditya Nalluri	USA	adityanalluri9@gmail.com	
ShongXian LEE	Singapore	shongxian5601@gmail.com	
Koffi Yao Lionel Stephen	Canada	yaolionelstephen.koffi@gmail.com	

Statement of integrity: By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an "X" above).

Team member 1	**Aditya Nalluri**
Team member 2	**Shongxian LEE**
Team member 3	**Koffi Yao Lionel Stephen**

Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.
Note: You may be required to provide proof of your outreach to non-contributing members upon request.

Contents

Autocorrelation

Definition: Technical definition using formulas or equations

Autocorrelations is to measure the degree of correlation of the variable between two-time intervals (Taylor, S. 2024). It is an issue that happens in regression model which can impact on the consistency and accuracy of the prediction. An assumption of linear regression is required to ensure the independency of the independent variables. If the assumption has broken, autocorrelation may result. The autocorrelation function has shown as below:

$$\rho_X(s,t) = \frac{\gamma_X(s,t)}{\sqrt{\gamma_X(s,s)\gamma_X(t,t)}}$$

It is used to measure the linear relationship between Xs and Xt. The value of the relationship is range between -1 and 1.

Description:

If the value of autocorrelation is not close to or equal to zero. The independent of the predictor variables is in doubt. Hence, it will affect the accuracy of the prediction. In the stock market, autocorrelation could be happened as the stock price will never be increased all the time and will be declining after sometime. In the equation of predicting the stock market, the stock prices in the past are also an indicator (independent variable) to forecast stock price's future value. Thus, in the prediction, autocorrelation should be avoided for better prediction which reduce the movement trend and biases from the past stock price.

Demonstration:

Date2	UST10Y	GOOGLE
2016-01-04	2.245	741.840027
2016-01-05	2.248	742.580017
2016-01-06	2.177	743.619995
2016-01-07	2.153	726.390015
2016-01-08	2.130	714.469971
...	...	...
2021-12-23	1.493	2942.850098
2021-12-27	1.481	2961.280029
2021-12-28	1.481	2928.959961
2021-12-29	1.543	2930.090088
2021-12-30	1.515	2920.050049

1505 rows × 2 columns

The data has selected US dollar index daily return and google stock price from 1st April 2016 to 30th December 2021. In the upcoming section, it will perform visualization for detecting autocorrelation by comparing different lag.

ARIMA model

ARIMA (p,d,q) model is extend from ARMA model which have autoregression model and moving average by adding order of differencing. By using this model, it can observe the existence of autocorrelation and also select the best fit model for prediction.

Google

```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-8091.969, Time=0.26 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=-8109.119, Time=0.11 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=-8108.303, Time=0.12 sec
 ARIMA(2,1,0)(0,0,0)[0]             : AIC=-8107.506, Time=0.39 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=-8107.472, Time=0.14 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=-8105.502, Time=0.18 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-8112.960, Time=0.14 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-8094.611, Time=0.14 sec
 ARIMA(2,1,0)(0,0,0)[0] intercept   : AIC=-8111.185, Time=1.46 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-8111.181, Time=0.32 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-8112.242, Time=0.48 sec
 ARIMA(2,1,1)(0,0,0)[0] intercept   : AIC=-8109.167, Time=2.12 sec

Best model:  ARIMA(1,1,0)(0,0,0)[0] intercept
Total fit time: 5.911 seconds
```

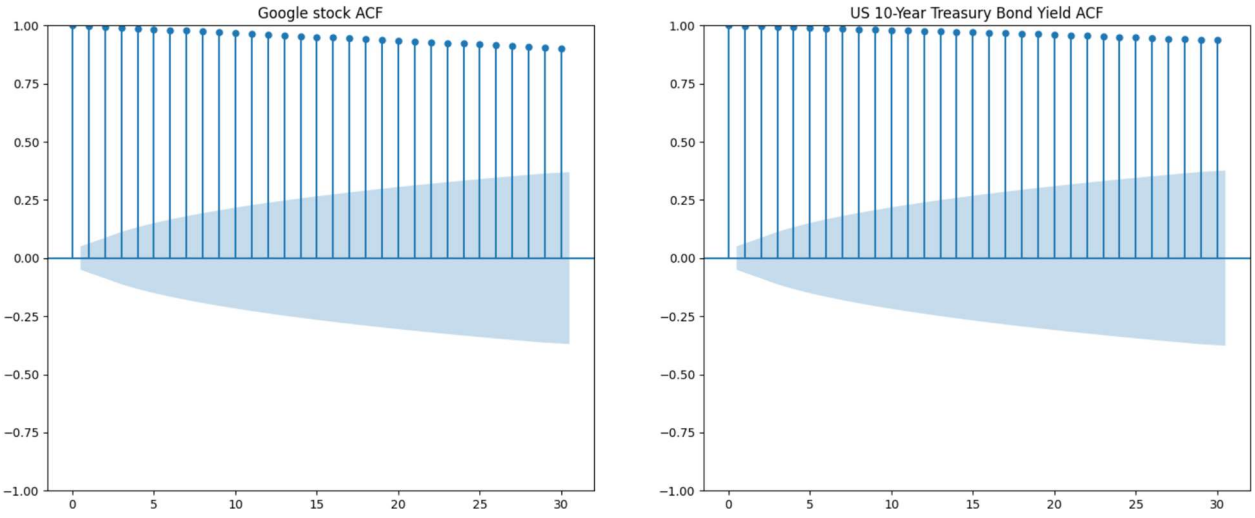
Through the ARIMA (p,d,q) model, the range from 0 to 3 have be used to test which combination is the most appropriate to measure the goodness of fit a model. The method that been used is Akaike's information criterion (AIC). The lower of AIC is prefer which mean it capable to balance the discrepancy between less precise predictions and number of parameters. From the picture, the best model is ARIMA (1,1,0) with -8,109.119

US 10-Year Treasury Bond

```
Performing stepwise search to minimize aic
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-5762.912, Time=0.08 sec
 ARIMA(1,1,0)(0,0,0)[0]             : AIC=-5761.366, Time=0.35 sec
 ARIMA(0,1,1)(0,0,0)[0]             : AIC=-5761.428, Time=1.07 sec
 ARIMA(1,1,1)(0,0,0)[0]             : AIC=-5772.009, Time=2.63 sec
 ARIMA(2,1,1)(0,0,0)[0]             : AIC=-5760.976, Time=1.97 sec
 ARIMA(1,1,2)(0,0,0)[0]             : AIC=-5775.288, Time=0.91 sec
 ARIMA(0,1,2)(0,0,0)[0]             : AIC=-5770.517, Time=0.57 sec
 ARIMA(2,1,2)(0,0,0)[0]             : AIC=-5810.901, Time=1.46 sec
 ARIMA(3,1,2)(0,0,0)[0]             : AIC=-5772.966, Time=1.53 sec
 ARIMA(2,1,3)(0,0,0)[0]             : AIC=-5830.018, Time=1.74 sec
 ARIMA(1,1,3)(0,0,0)[0]             : AIC=-5774.404, Time=1.47 sec
 ARIMA(3,1,3)(0,0,0)[0]             : AIC=-5828.165, Time=3.06 sec
 ARIMA(2,1,3)(0,0,0)[0] intercept   : AIC=-5827.771, Time=3.68 sec

Best model:  ARIMA(2,1,3)(0,0,0)[0]
Total fit time: 20.566 seconds
```

Based on the prediction for US 10-Year Treasury Bond, the best fit model is ARIMA (2,1,3) with -5,830.018.

Diagram & Diagnosis

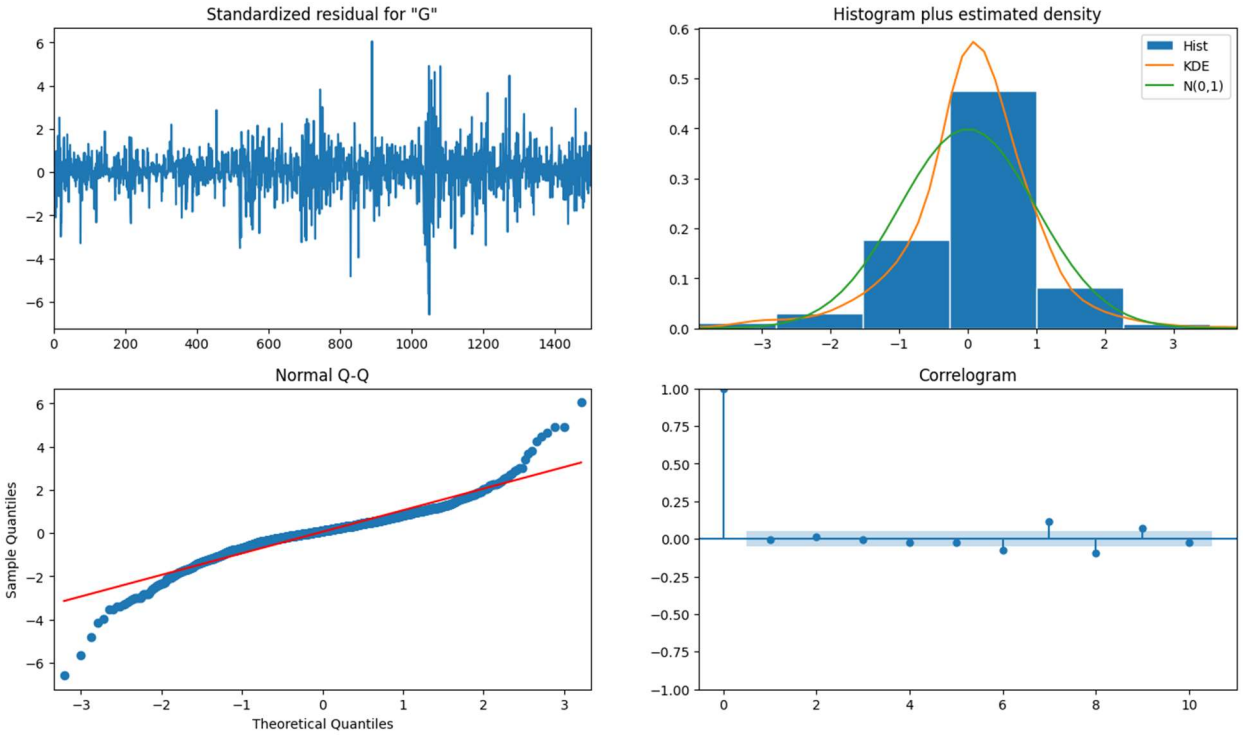


This is the diagram to visualize if the autocorrelation exists in the dataset.

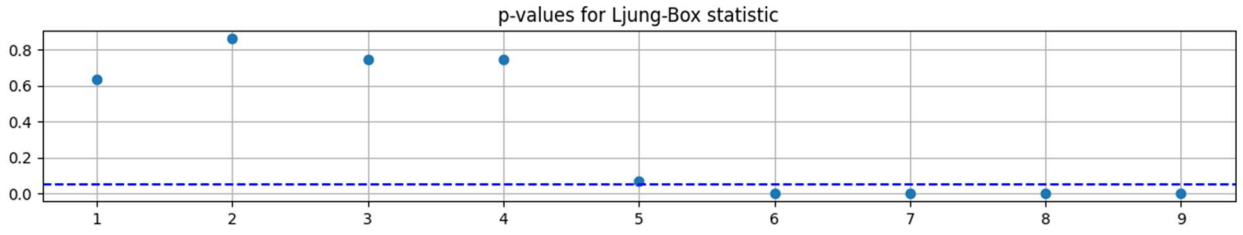
To examine the degree of autocorrelation between data, it can observe with different lags. From two visualization, the value of correlation is 1 at 0 lag at the left side. It is because its correlation to itself initially, hence, it results with 1. Through the plot, there have shaded area with 95% confidence intervals around 0. If the data point does not go beyond the shaded area, it means there is no autocorrelation exist within the dataset. In fact, all the data points were exceeding the shaded area at all time lag. It can conclude that autocorrelation is significant.

The trend of both, google stock and US 10-year treasury bond yield were showing decreasing autocorrelation.

ARIMA model for US 10-Year Treasury Bond, ARIMA (1,1,0)



The graph that in the top left (standardized residual graph), it showed the data points are volatile at the mean value of 0. Additionally, it gets more volatile at the value between 1000 and 1200. With the white noise's assumption of mean value is 0 and variance is equal to one, hence, stationary able to prove. From the normal Q-Q plot, the theoretical quantities between -2 and 2.5, the data plots were showing a normality, but x < -2 and 2.5>x were showing non-normality as it far from the linear regression line. At the correlogram, lag 6 to lag 9 are significant.

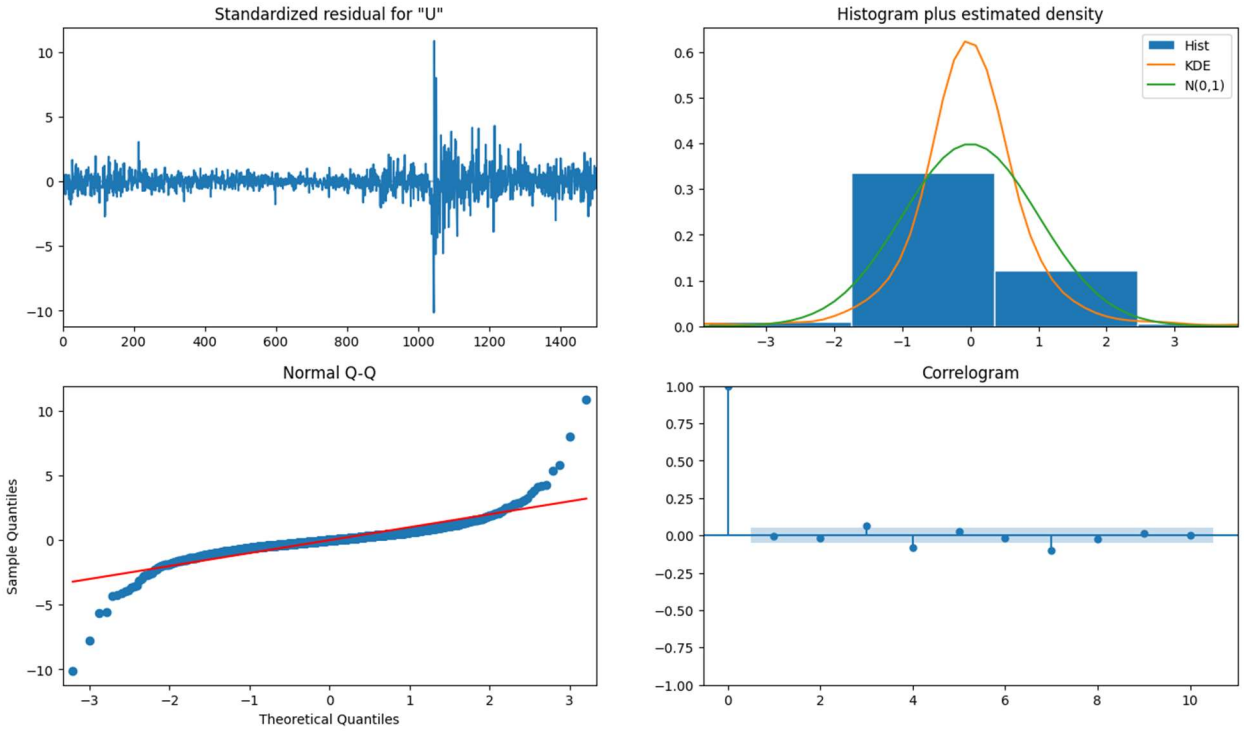


H0 (Null hypothesis): The residuals are independent as the correlation is small enough

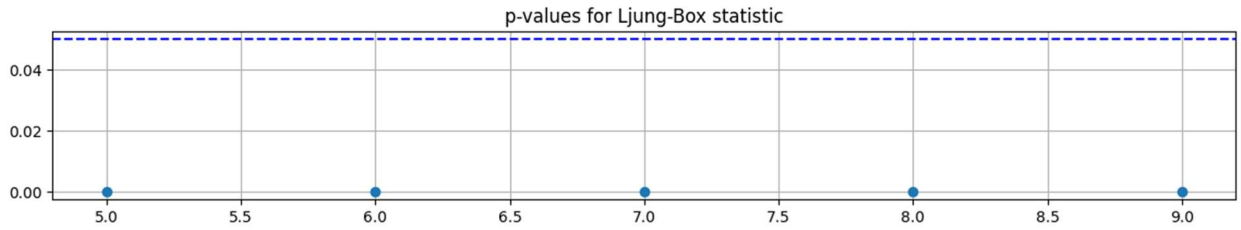
H1 (Alternative hypothesis): The residuals are not independent as the correlation is not small enough.

The lag from 1 to 5 is greater than 0.05 which fail to reject null hypothesis. Hence, it is independent as the autocorrelation is small enough in the group. The p-value from lag 6 to lag 9 is less than 0.05, hence, it can conclude that residuals are not independent as correlation is not small enough

ARIMA model for US 10-Year Treasury Bond, ARIMA (2,1,3)



The standardized residual is volatile start from approximately 1050. Based on the normal Q-Q plot, the tail and head are experiencing an extreme outlier. From the correlogram, lag 3, 4, and 7 are significant. The rest of the lags were closed to zero.



H0 (Null hypothesis): The residuals are independent as the correlation is small enough

H1 (Alternative hypothesis): The residuals are not independent as the correlation is not small enough.

As show in Ljung-statistics, the p-value is equal to 0 which less than 0.05, hence, it rejects null hypothesis. This refers that the residuals are not independent and indicate strong autocorrelation between different lag within the time intervals.

Damage:

In an ordinary least square regression model, it can have different independent variables to predict an output. The equation showed as below is the ordinary least square regression model.

$$Y=\beta 0+\beta 1X1+\beta 2X2+\beta 3X3+e$$

There have few assumptions that no multicollinearity, independent error terms, normal distributed of error terms.

If the independent of variables are unable to ensure, the output of the prediction, y^ will become inaccurate. It is because independent variables can affect each other.

From the visualization above US 10-Year Treasury Bond's data have autocorrelation among each other. Hence, the prediction may not be accurate and the modelling equation is not appropriate to use. It should come out with a solution to address the issue of autocorrelation.

Directions:

First of all, ARIMA model is suggested to determine the best fit model with the moving average, order of differencing, and autoregressive model.

Secondly, data transformation of square, square root, and log. This able to reduce the autocorrelation of the data before fitting into the linear regression model. It able to reduce the variance and skewness with the objective of normality of the distribution. The other transformation can refer to differencing method which able to remove the trend.

If the data is large enough and highly correlated, principal component analysis (PCA) can be adopted. PCA will create a new variable/ principal component that do not have any relationship with other principal component. This enhances the normal distribution and reduce the issue the multicollinearity.

Heteroskedasticity (error term are not constant)

Definition: Technical definition using formulas or equations

$$y1=\beta 0+\beta 1X11+\beta 2X21+\beta 3X31+e1$$

$$y2=\beta 0+\beta 1X12+\beta 2X22+\beta 3X32+e2$$

The equation stated above is the ordinary least square regression (OLS). The assumption of OLS are linear relationship, no multicollinearity, normal distributed and independent of error terms, and expected values of error terms are equal to 0.

The opposition of heteroskedasticity is homoskedasticity. The assumption of homoskedasticity is the error term between observations from dataset are independent which is constant. A constant error terms are independent and identically distributed. The equation of error terms is stated as below:

$$e\sim N\ (0,\ \sigma 2)$$

The mean value is equal to 0 with a constant variance.

If the error terms are not constant, it is heteroskedastic

Description:

If heteroskedasticity exists, the residuals between measured values of the dependent variables and predicted values of dependent value will result a cone shape of distribution. The wider the cone shape, the volatile of the residual values.

To determine the existence of normality of the residuals scatter:

H0: Homoscedasticity exists

H1: Heteroskedasticity exists

If the p-value is less than 0.05, null hypothesis (H0) will be rejected and heteroskedasticity exists.

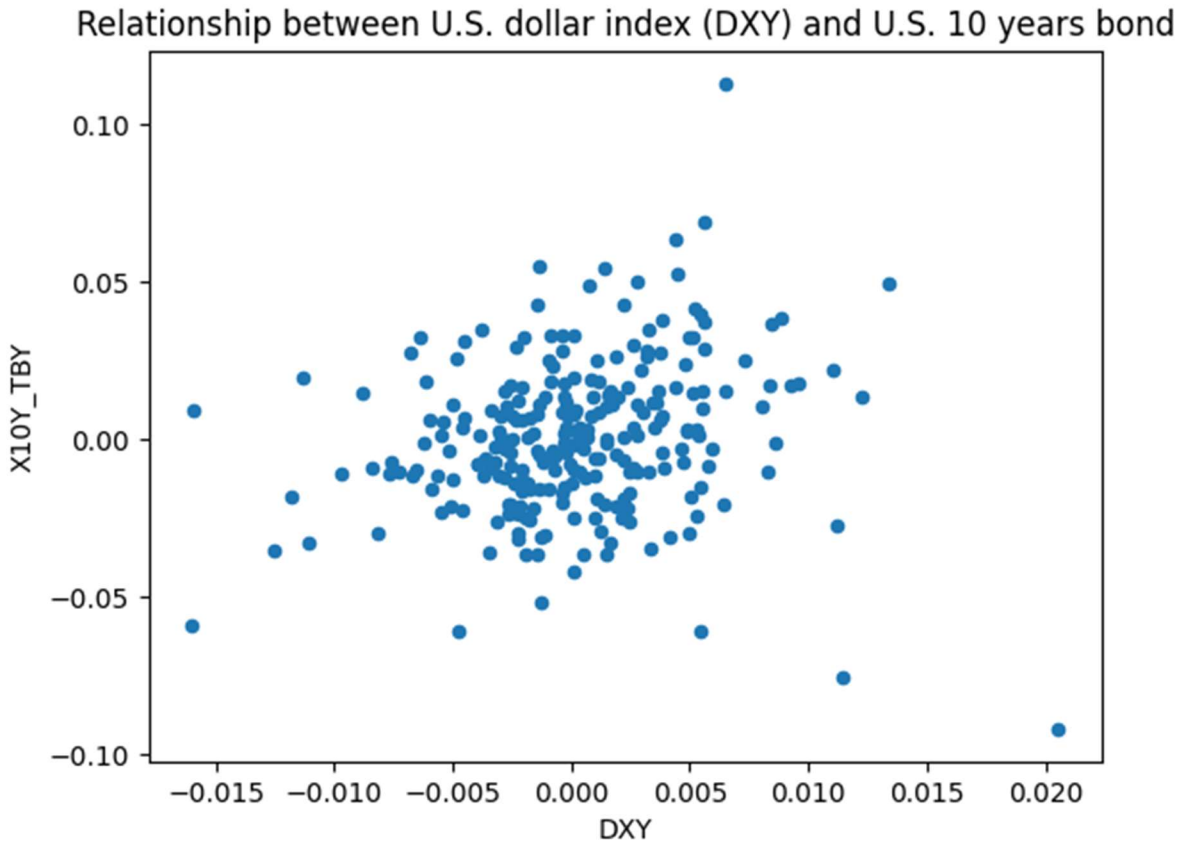
1. Demonstration:

	Unnamed: 0	Date	DXY	METALS	OIL	US_STK	INTL_STK	X13W_TB	X10Y_TBY	EURUSD	YEAR
0	1	1/4/2016	0.002433	0.024283	-0.007559	-0.013980	-0.019802	0.047297	-0.010577	-0.007316	2016
1	2	1/5/2016	0.005361	-0.004741	-0.021491	0.001691	-0.001263	0.322581	0.001336	-0.002436	2016
2	3	1/6/2016	-0.002213	0.013642	-0.055602	-0.012614	-0.015171	0.000000	-0.031584	-0.006978	2016
3	4	1/7/2016	-0.009679	0.035249	-0.020606	-0.023992	-0.019255	-0.073171	-0.011024	0.002512	2016
4	5	1/8/2016	0.003258	-0.028064	-0.003306	-0.010977	-0.010471	0.000000	-0.010683	0.013636	2016
...	...	...	...	...	...	...	...	...	...	...	...
245	246	12/23/2016	-0.000776	0.016432	0.001322	0.001464	0.000818	0.014199	-0.003917	0.000866	2016
246	247	12/27/2016	0.000097	0.023845	0.016598	0.002481	0.000409	-0.010000	0.007865	0.001851	2016
247	248	12/28/2016	0.002330	0.017105	0.002968	-0.008265	-0.001226	0.000000	-0.022240	0.000513	2016
248	249	12/29/2016	-0.005617	0.061320	-0.005364	-0.000223	0.005728	-0.111111	-0.011572	-0.004334	2016
249	250	12/30/2016	-0.002824	-0.038762	-0.000930	-0.003655	0.002034	0.090909	-0.012515	0.015197	2016

250 rows × 11 columns

The data using is provided by WorldQuant University. The data of U.S. dollar index (DXY) and U.S. 10 years bond. The data was collected from 4th January 2016 to 30th December 2016.

Diagram & Diagnosis



From the scatterplot, it provides the visualization when DXY increase, it also led X10Y_TBY increase. After the value 0.005 of DXY, the value is widely distributed and uncertain. It can observe that there has a cone shape and indicate heteroskedasticity.

OLS Regression Results

Dep. Variable:	DXY	**R-squared:**	0.030
Model:	OLS	**Adj. R-squared:**	0.026
Method:	Least Squares	**F-statistic:**	7.738
Date:	Tue, 26 Mar 2024	**Prob (F-statistic):**	0.00582
Time:	14:50:56	**Log-Likelihood:**	986.78
No. Observations:	250	**AIC:**	-1970.
Df Residuals:	248	**BIC:**	-1963.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0001	0.000	0.474	0.636	-0.000	0.001
X10Y_TBY	0.0336	0.012	2.782	0.006	0.010	0.057

Omnibus:	26.741	**Durbin-Watson:**	1.909
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	94.918
Skew:	0.322	**Prob(JB):**	2.45e-21
Kurtosis:	5.949	**Cond. No.**	40.7

H0: Homoscedasticity exists

H1: Heteroskedasticity exists

$\hat{Y}$ (DXY) = 0.0001 + 0.0336 *B* (X10Y_TBY)

It has shown the OLS Regression result, the positive correlation between DXY & X10Y_TBY has showed with 0.0336 which mean X10Y_TBY increase by 1 percent, DXY will be increased by 0.0336.

Referring to kurtosis of 3 means it is a normal distribution, with the information of 5.949 and deduct 3, positive 2.949, it indicates leptokurtic with higher peak and thin tails. However, with a skewness of 0.322 and right skewed distribution has result. Hence, it is not a normal distribution.

Lagrange multiplier statistic	13.897238
p-value	0.000193
f-value	14.597520
f p-value	0.000168

Breusch-Pagan test is used to address the existence of Heteroskedasticity. With the p-value of 0.000193, it rejects the null hypothesis, hence, hheteroskedasticity exists.

Damage:

The impact of heteroscedasticity can result the information of standard errors and confidence intervals are not reliable for further analysis. It also can lead to overestimate or underestimate of the result. However, it also does not follow the assumption of OLS.

Since the residuals values are not constant. Thus, it can result volatile and increase the standard error. The estimation of the model can be weak which refer to the R-squared. The issue of autocorrelation can be result as the error terms are not independent in the dataset (Rehal, V. 2023).

Directions:

Data transformation is able to solve Heteroscedasticity issue. The common methods are log, square roots and square which able reduce the possibility of correlation. Second method, is reduce the number by using the rate of change instead of the actual amount.

The last method is weight least square that added the absolute residuals and fitted values. This able to improve the prediction of the model. If weight least square is unable to fix the issue, robust regression is recommended to reduce the impact of outliers.

Over-reliance on Normality (e.g., the Gaussian Distribution)

Definition:

The Gaussian distribution, also known as the normal distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. The formula for the probability density function of a Gaussian distribution is given by:
Normal Distribution Formula: [1]

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ = Mean
σ = Standard deviation
x = Normal random variable

Description:

In financial time series analysis, there is a common assumption that returns, or other financial data follow a Gaussian distribution. This assumption facilitates model building, statistical testing, and theoretical analysis. However, real-world financial data often exhibit properties such as fat tails (higher likelihood of extreme outcomes than predicted by the Gaussian distribution) and skewness, making the over-reliance on normality problematic. It can lead to underestimation of the probability of extreme events, mispricing of risk, and inadequate risk management strategies. Addressing over-reliance on normality involves employing models that can accommodate non-normal distributions, such as the use of heavy-tailed distributions, or applying transformations to data to mitigate the effects of non-normality.

Demonstration:

Our analysis of NVDA stock data, spanning from April 13, 2017, to March 12, 2024, reveals several key insights into the stock's behavior, leveraging a combination of histogram, Q-Q plot, volatility clustering, and time series analysis.

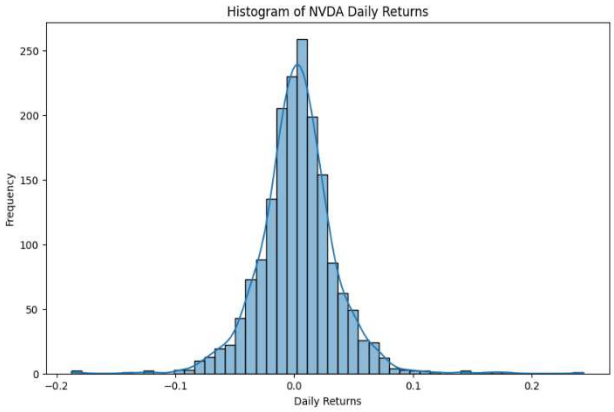
Histogram of Daily Returns: The histogram presents a very slight asymmetry in daily returns, indicating a mild deviation from a perfectly normal distribution. This slight skewness suggests that while NVDA's stock returns are broadly symmetrical, there are subtle biases in how gains and losses are distributed, potentially hinting at more frequent small gains over losses or vice versa.

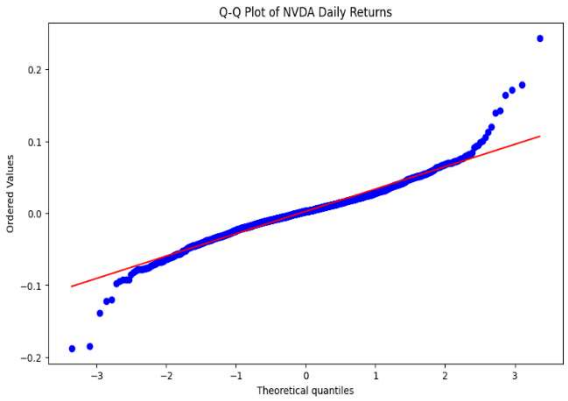
Q-Q Plot Analysis: The Q-Q plot further elucidates the deviation from normality, particularly highlighting extreme outliers in the tails and head of the distribution. These outliers signify that extreme price movements in NVDA's stock, both positive and negative, occur more frequently than what would be expected under a normal distribution. This pattern underscores the presence of fat tails, a common characteristic in financial time series that points to a higher likelihood of experiencing significant market movements.

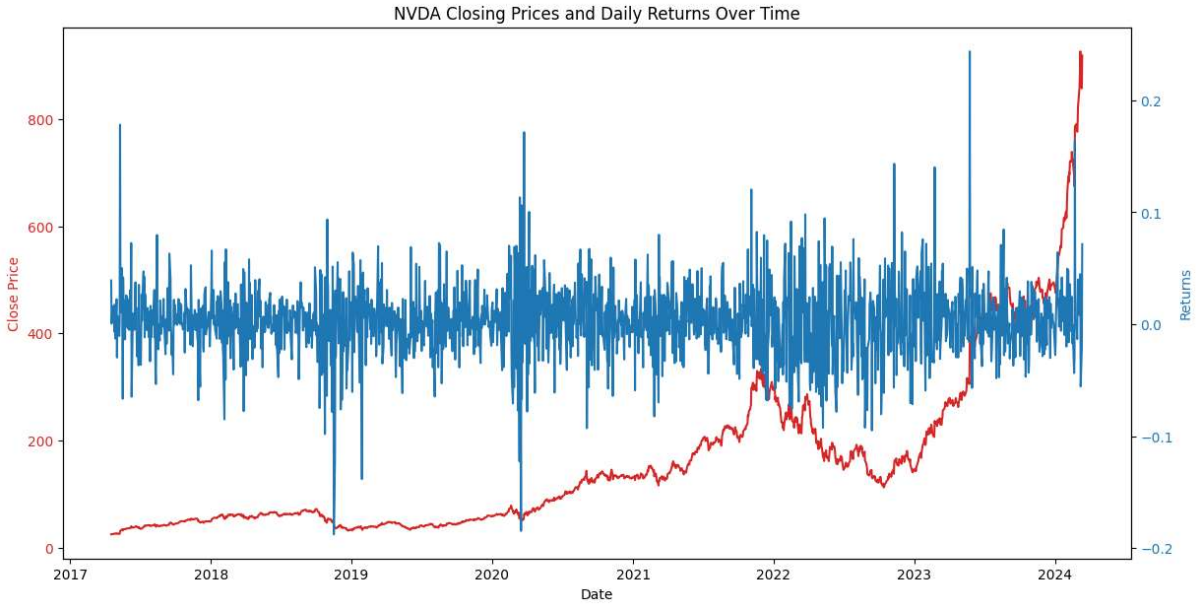
Volatility Clustering: Our examination of squared returns illuminates periods of pronounced volatility clustering. This observation indicates that NVDA's stock experiences phases of heightened volatility, particularly around significant events such as quarterly earnings reports. Such patterns are indicative of the market's reaction to new information, where investors adjust their positions based on the company's financial health and future outlook, leading to increased price fluctuations.

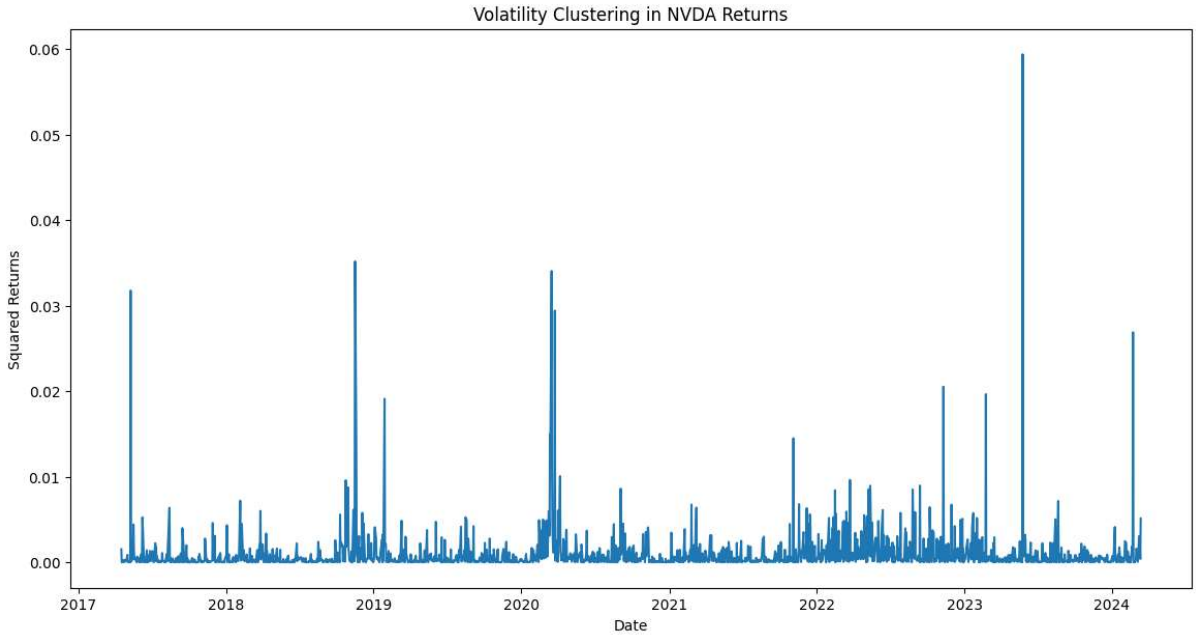
Time Series Plot of Closing Prices and Returns: The time series analysis provides a clear visualization of NVDA's stock price and returns over time. Initially, the stock price consolidates, followed by a period of exponential growth. This growth trajectory is likely driven by increasing demand for GPUs and advancements in AI technology, reflecting Nvidia's strengthening position in these sectors. The accompanying returns are notably volatile, fluctuating in response to market trends, company developments, and broader economic factors.

```
         Date       Open       High        Low      Close  Adj Close    Volume  \
1  2017-04-17  24.002501  24.809999  23.950001  24.807501  24.480764  49729200
2  2017-04-18  24.662500  24.885000  24.400000  24.822500  24.495569  37011600
3  2017-04-19  25.000000  25.245001  24.852501  24.920000  24.591782  38082800
4  2017-04-20  25.067499  25.362499  24.852501  25.315001  24.981577  40401600
5  2017-04-21  25.209999  25.447500  25.090000  25.420000  25.085196  34189600

    Returns
1  0.039166
2  0.000605
3  0.003928
4  0.015851
5  0.004148
                Open         High          Low        Close    Adj Close  \
count  1738.000000  1738.000000  1738.000000  1738.000000  1738.000000
mean    160.203268   163.133229   157.202882   160.345772   160.046793
std     149.777525   152.479675   146.991536   149.974090   150.069824
min      24.002501    24.809999    23.950001    24.807501    24.480764
25%      52.520624    53.036875    51.354374    52.403750    52.127521
50%     124.882503   126.939999   121.015003   124.671254   124.503120
75%     209.632504   212.930000   206.507496   208.272500   207.941250
max     951.380005   974.000000   896.020020   926.690002   926.690002

             Volume      Returns
count  1.738000e+03  1738.000000
mean   5.006017e+07     0.002608
std    2.498636e+07     0.031858
min    9.788400e+06    -0.187559
25%    3.477960e+07    -0.014030
50%    4.538435e+07     0.002893
75%    5.967272e+07     0.019257
max    3.692928e+08     0.243696
```









Shapiro-Wilk Test Statistic: 0.952, P-Value: 1.90e-23, Skewness: 0.245, Kurtosis: 5.28

Collectively, these analyses paint a picture of NVDA's stock as one characterized by slight asymmetry in returns, more frequent extreme movements than a normal distribution would predict, periods of significant volatility, and robust growth fueled by technological advancements. The presence of fat tails and volatility clustering, in particular, challenge the assumption of normality and constant volatility in financial time series analysis, underscoring the need for models that can accommodate these complexities. Our findings highlight the dynamic

nature of the stock market, where investor sentiment, company performance, and external events interplay to shape the behavior of stock prices.

Diagnosis: Recognizing the Problem

In our analysis of NVDA stock returns, we delved into a blend of statistical tests and visual diagnostics, including the Shapiro-Wilk test, skewness, kurtosis, histograms, Q-Q plots, time series plots, and observations of volatility clustering. The Shapiro-Wilk test's P-value of approximately 1.90e-23 compellingly rejects the normality assumption, while a kurtosis of 5.28 signifies pronounced fat tails, and a skewness of 0.245 suggests slight asymmetry. These findings are visually corroborated by the Q-Q plot, which highlights outliers at both ends, indicating extreme values are more prevalent than normality would predict.

Histogram analysis further reveals this slight asymmetry in daily returns, and the time series plot shows NVDA's stock price experiencing significant growth, punctuated by periods of volatility, as seen in the volatility clustering analysis. These periods likely correspond to market reactions to various events, showcasing the stock's dynamic response beyond what a stable, normal distribution could encapsulate.

This comprehensive approach underlines the inadequacy of relying solely on normality in financial time series analysis, especially for NVDA stock returns. The observed skewness, fat tails, and volatility clustering suggest that models assuming normal distribution may underestimate the likelihood and impact of extreme market movements. Consequently, this analysis prompts the exploration of more sophisticated models that can better accommodate the empirical characteristics of financial data, leading to more accurate risk assessments and investment strategies.

Damage: The Implications of Over-Reliance on Normality

The consequences of assuming normality in financial data, such as stock returns, can be profound. Normality implies a symmetrical distribution of data around the mean, with tail events (extremely high or low returns) considered highly unlikely. However, our analysis reveals that NVDA's returns exhibit fat tails, as indicated by the high kurtosis value. This means extreme price movements are more common than a normal distribution would suggest.

The damage from over-reliance on normality primarily comes from underestimating the risk of extreme market movements. Traditional risk management models, like the Value at Risk (VaR) calculated under normality assumptions, might provide a false sense of security. In reality, the likelihood of experiencing losses that exceed these risk measures can be significantly higher than predicted, leading to potential financial distress or systemic risks in the broader market.

Directions: Moving Beyond Normality

Given the diagnostic evidence and the potential for misestimating risk, it's crucial to explore models that better accommodate the actual behavior of financial returns. Here are some directions:

1. Non-Parametric Models: These models do not assume a specific probability distribution, making them more flexible in capturing the characteristics of financial data. Examples include historical simulation for VaR calculations.
2. Heavy-Tailed Distributions: Models that assume heavy-tailed distributions, such as the Student's t-distribution, can more accurately capture the likelihood of extreme returns. Implementing these distributions in risk management frameworks can provide a more realistic assessment of tail risks.
3. GARCH Models: The Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model is particularly effective for modeling financial time series data that exhibits time-varying volatility of volatility clustering, a feature our analysis suggests might be present.
4. Transformations and Robust Estimators: Applying transformations to the data, such as logarithmic returns, or using robust statistical estimators can also mitigate the impact of non-normality and provide more reliable insights.

In conclusion, while the assumption of normality offers simplicity and mathematical convenience, our analysis underscores the importance of acknowledging and addressing its limitations. By employing alternative models and methods that more accurately reflect the empirical behavior of financial data, we can enhance our understanding of market dynamics and improve our risk management practices.

Modeling Randomness

Technical definition

In econometrics, we aim to model economic phenomena by considering both deterministic variables and an unobserved error term. The unobserved error term, often referred to as white noise in time series analysis, represents the stochastic component of the equation, capturing the randomness or unforeseen factors affecting the relationship between variables. Various models are employed to characterize this error term, each tailored to different aspects of the data's behavior.

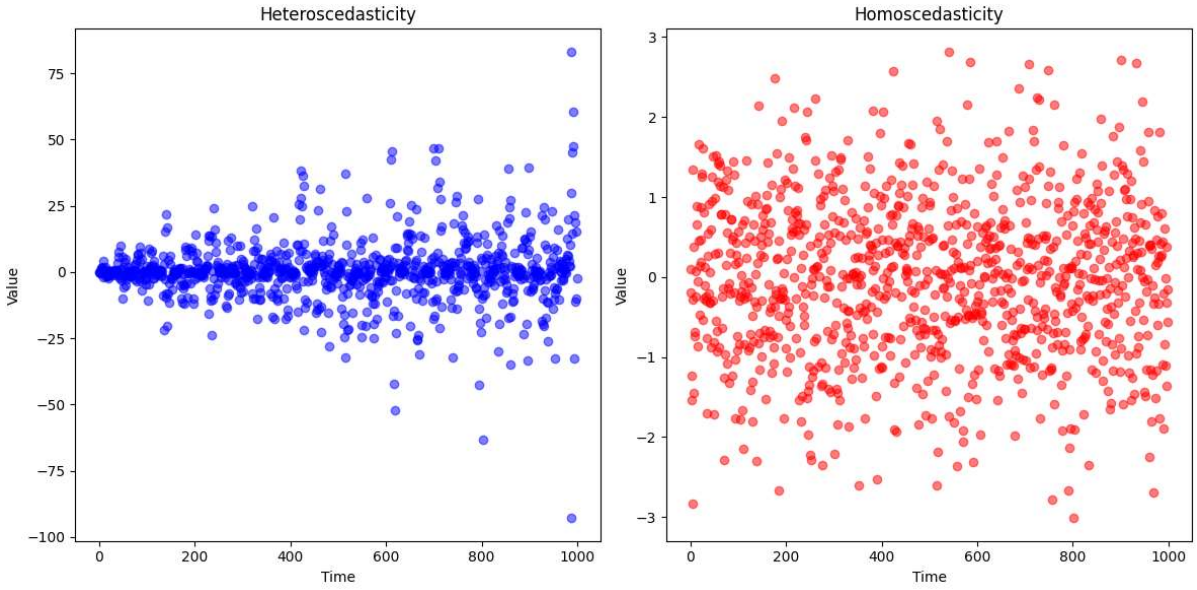
Among these models are Autoregressive (AR), Moving Average (MA), Generalized Autoregressive Conditional Heteroskedasticity (GARCH), and Exponential GARCH (EGARCH). These models are particularly useful for capturing the volatility clustering phenomenon observed in financial markets, where periods of high volatility tend to cluster together.

Let's delve into one of the commonly used methods for modeling the error term in the context of stock volatility: the GARCH model. GARCH (p, q) model is a model which, the error terms, can be split into a stochastic piece $'z_t$ and a time-dependent standard deviation σ_t' characterizing the typical size of the terms so that $\varepsilon_t = \sigma_t z_t$. The random variable is a strong white noise process while it is an ARMA process. is a constant and are the coefficients

$$\sigma_i^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{p} \beta_i \sigma_{t-i}^2 \P$$

Demonstration and Visualization:

The presence of randomness often affects the error term in equations. One important phenomenon to consider is heteroscedasticity, which occurs when the variance of the error term is not constant across different levels of the independent variables. To illustrate this concept, we simulate two types of data: heteroscedasticity and homoscedasticity. Heteroscedasticity, characterized by varying variance, is showcased through a simulated time series dataset comprising 1000 observations. This data exhibits a 'corn-like' graph pattern, demonstrating how the variance of errors changes with the independent variable. On the other hand, homoscedasticity refers to a scenario where the variance of errors remains constant. We also simulate homoscedastic data to provide a clear comparison. Below are two charts displaying these simulated datasets, illustrating the distinction between heteroscedastic and homoscedastic phenomena.



Demonstration:

modeling randomness is essential for capturing the inherent uncertainty and variability in economic data. Here are some methodologies that has been used to do so:

- **Stochastic Processes**: are mathematical models used to represent the evolution of variables over time in a probabilistic manner. Common stochastic processes used in econometrics include random walks, autoregressive processes (AR), moving average processes (MA), autoregressive integrated moving average processes (ARIMA), and more advanced models like GARCH for modeling time-varying volatility.
- **Monte Carlo Simulations**: involve generating random samples from probability distributions to estimate the behavior of complex systems. In econometrics, Monte Carlo simulations are often used to evaluate the properties of estimators, test statistics, and forecasting models under various assumptions and scenarios.
- **Bootstrapping**: Bootstrapping is a resampling technique used to estimate the sampling distribution of a statistic by repeatedly sampling from the observed data with replacement. Bootstrapping is particularly useful when the underlying distribution of the data is unknown or difficult to model parametrically.
- **Bayesian Econometrics**: Bayesian econometrics is an approach that incorporates Bayesian probability theory into econometric analysis. It allows for the estimation of parameters by updating prior beliefs based on observed data, providing a framework for modeling uncertainty and making probabilistic inferences.
- **Simulation-Based Estimation**: Simulation-based estimation techniques, such as the method of moments, maximum likelihood estimation, and Bayesian methods, involve

generating simulated data based on a specified model and comparing it with the observed data to estimate model parameters.

- **Robust Estimation**: Econometric models often encounter heteroscedasticity, where the variance of the errors is not constant across observations. Techniques such as robust standard errors, weighted least squares, and generalized least squares are employed to account for heteroscedasticity and improve the reliability of parameter estimates.

These techniques, among others, enable economists and econometricians to effectively model and analyze the randomness inherent in economic data, providing insights into economic phenomena, forecasting future trends, and making informed policy decisions.

Damage:

All models work with their own assumption and hypothesis related to the error term. Generally there are five assumptions that should be met to better model the randomness of the model. Those assumptions are stationarity, Independence, normality, homoscedasticity, and serial independence. Not comply with one of those assumptions can lead to various consequences as it's mentioned here:

- **Biased estimates:** If the randomness in the data is not appropriately captured by the model, the estimates of the model parameters may be biased. This can result in inaccurate predictions or conclusions drawn from the model.
- **Inefficient inference:** The estimated standard errors of the parameters may be incorrect, which affects the precision of the estimates and can lead to wider confidence intervals.
- **Incorrect hypothesis testing:** Test statistics may be biased, leading to incorrect conclusions about the significance of variables or the overall fit of the model.
- **Invalid predictions:** predictions made by the model may be unreliable. This can lead to poor decision-making and ineffective planning based on the model's outputs.

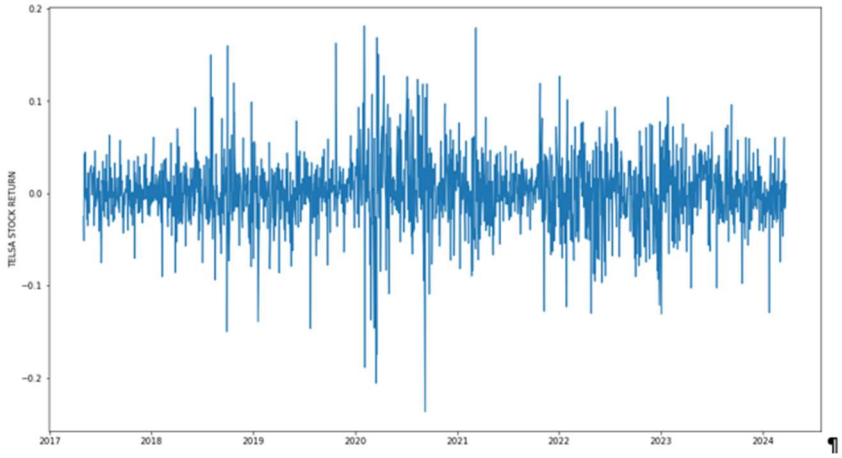
Mis-modeling randomness can lead to violations of these assumptions, compromising the validity of the model and the interpretations drawn from it.

Direction:

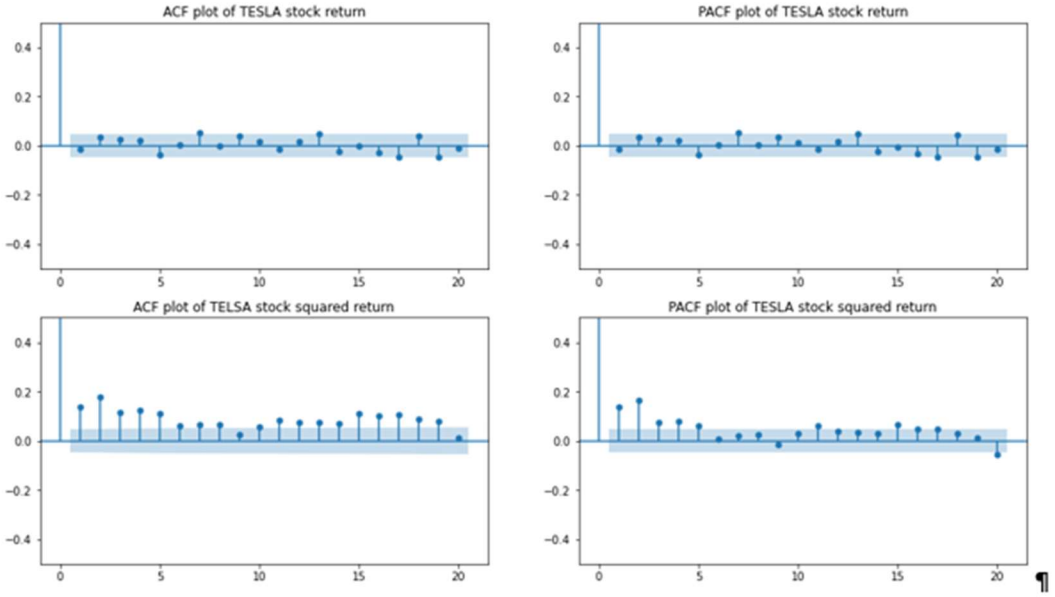
There are several ways to handle problems related to model misspecification of random error. those are the following step:there are several steps one can take to address them effectively:

- **Visualization**: Utilize various graphical tools such as scatter plots, Q-Q plots, Partial Autocorrelation Function (PACF), and Autocorrelation Function (ACF) charts. These visualizations can help in assessing the randomness and identifying any patterns or deviations from the expected behavior.
- **Statistical Tests**: Employ a range of statistical tests to ensure that the randomness conforms to the model assumptions. Some commonly used tests include:
    - Stationarity Test: Assess whether the time series data exhibits stationarity, which is often crucial for time series modeling.
    - Normality Test: Check for the normality of the error term distribution. Tests like the Shapiro-Wilk test or the Jarque-Bera test can be employed for this purpose.
    - Heteroscedasticity Test: Detect whether the variance of the error terms varies systematically across observations. Tests such as the Breusch-Pagan test can be useful in identifying heteroscedasticity.
- **Econometric Modeling**: Consider employing more advanced econometric modeling techniques to better capture the underlying structure of the data and mitigate the effects of model mis-specification like AR, MA, ARMA, ARIMA, GARCH etc.

We have analyzed the returns of Tesla stocks from 2017 to 2024 to assess their behavior. Upon visual inspection, it's evident that the returns of Tesla stock exhibit volatility clustering. This means that periods of high volatility tend to cluster together, while periods of low volatility also cluster together. Understanding this volatility clustering phenomenon is crucial for interpreting market dynamics, informing trading strategies, and managing risks effectively.



After also look into the ACF and PACF plot, It seems likes there is not lag in the model but there is a volatility clustering.



One approach to address volatility clustering is to test for heteroscedasticity. The results of the Ljung-Box and Box-Pierce tests indicate that the residuals exhibit heteroscedasticity.

Ljung-Box and Box-Pierce tests on standardized residuals

	lb_stat	lb_pvalue	bp_stat	bp_pvalue
1	0.015749	0.900130	0.015722	0.900216
2	1.789912	0.408626	1.785799	0.409467
3	2.198199	0.532303	2.192911	0.533346
4	3.938766	0.414356	3.927465	0.415911
5	4.172965	0.524792	4.160720	0.526516
6	6.807447	0.339022	6.783070	0.341376
7	13.391747	0.063120	13.333254	0.064393
8	13.637312	0.091722	13.577405	0.093467
9	14.848833	0.095173	14.781254	0.097122
10	14.990041	0.132425	14.921486	0.134951

Ljung-Box and Box-Pierce tests on stanrdized squared residuals

	lb_stat	lb_pvalue	bp_stat	bp_pvalue
1	0.599913	0.438611	0.598877	0.439007
2	3.257629	0.196162	3.250472	0.196865
3	3.308802	0.346419	3.301498	0.347434
4	3.319218	0.505893	3.311878	0.507052
5	3.331314	0.649051	3.323925	0.650181
6	4.666775	0.587205	4.653236	0.588993
7	5.557267	0.592286	5.539114	0.594470
8	5.783120	0.671511	5.763667	0.673685
9	6.482995	0.690772	6.459110	0.693233

10 7.965190 0.632237 7.931065 0.635570

ARCH LM test for conditional heteroskedasticity

ARCH-LM Test

H0: Standardized residuals are homoskedastic.

ARCH-LM Test

H1: Standardized residuals are conditionally heteroskedastic.

Statistic: 20.6267

P-value: 0.7132

Distributed: chi2(25)

Volatility Model:

- **OMEGA** (Intercept): The estimated coefficient is 0.1453 with a standard error of 0.01388. The t-statistic is 10.467, indicating that the intercept is statistically significant at a very high level (p-value < 0.001). The 95% confidence interval for **omega** ranges from 0.118 to 0.172.
- **alpha[1]** (ARCH Coefficient): The estimated coefficient is 0.1463 with a standard error of 0.04659. The t-statistic is 3.140, indicating that the coefficient is statistically significant at the 0.05 level (p-value = 0.00169). The 95% confidence interval for **alpha[1]** ranges from 0.055 to 0.238.

Distribution:

- **nu** (Degrees of Freedom): The estimated coefficient is 3.4030 with a standard error of 0.286. The t-statistic is 11.880, indicating that the parameter is statistically significant at a very high level (p-value < 0.001). The 95% confidence interval for **nu** ranges from 2.842 to 3.964.

These coefficients are typical in ARCH models with a Student's t distribution, where **OMEGA** represents the constant term,**alpha[1]** is the autoregressive coefficient for the first lag of squared residuals (ARCH effect), and **nu** is the degrees of freedom parameter for the t-distribution, controlling for the heavy tails often observed in financial data.

Overall, the ARCH model with a Student's t distribution appears to provide a statistically significant fit to the volatility of the data, with both the intercept and ARCH coefficient being significant. The degrees of freedom parameter (**nu**) suggests that the residuals may have heavier tails than a normal distribution.

Zero Mean - ARCH Model Results

Dep. Variable:	rets_TSLA	**R-squared:**	0.000
Mean Model:	Zero Mean	**Adj. R-squared:**	0.001
Vol Model:	ARCH	**Log-Likelihood:**	-657.006
Distribution:	Standardized Student's t	**AIC:**	1320.01
Method:	Maximum Likelihood	**BIC:**	1336.39
		No. Observations:	1735
Date:	Tue, Mar 26 2024	**Df Residuals:**	1735
Time:	08:59:48	**Df Model:**	0

Volatility Model

	coef	std err	t	P>|t|	95.0% Conf. Int.
omega	0.1453	1.388e-02	10.467	1.219e-25	[0.118, 0.172]
alpha[1]	0.1463	4.659e-02	3.140	1.691e-03	[5.497e-02, 0.238]

Distribution

	coef	std err	t	P>|t|	95.0% Conf. Int.
nu	3.4030	0.286	11.880	1.505e-32	[2.842, 3.964]

Reference

"Autocorrelation". *Corporate Finance Institute (CFI),* uploaded by Taylor, S., *2024*, https://corporatefinanceinstitute.com/resources/data-science/autocorrelation/. Accessed 25th March 2024.

"Heteroscedasticity: Causes and Consequences" *Spur Economics*, uploaded by Rehal, V., 8th February 2023, https://spureconomics.com/heteroscedasticity-causes-and-consequences/. Accessed 27th March 2024.